# Designing Repeat Proteins: A Modular Approach to Protein Design

Fabio Parmeggiani[1] and Po-Ssu Huang[2]

1. School of Chemistry, School of Biochemistry, University of Bristol, Bristol, UK
2. Bioengineering Department, Stanford University, Stanford, California, USA

Correspondence to: fabio.parmeggiani@bristol.ac.uk or possu@stanford.edu

## Abstract

Repeat proteins present unique opportunities for engineering because of their modular nature that potentially allows LEGO® like construction of macromolecules. Nature takes advantage of these properties and uses this type of scaffold for recognition, structure, and even signaling purposes. In recent years, new protein modeling tools facilitated the design of novel repeat proteins, creating possibilities beyond naturally occurring scaffolds alone. We highlight here the different design strategies and summarize the various structural families and novel proteins achieved.

## Introduction

A goal of protein engineers is to understand how sequence and structure features contribute to the makeup of proteins. However, polypeptide complexity and variability complicate the analysis of these features. Proteins that carry distinct patterns of repetitive sequences (and therefore structural features) are ideal systems with reduced complexity to inform our understanding of proteins. These repeat proteins are widespread in nature [1] and the evolutionary process that created them is quite remarkable: a segment of sequence that is structurally compatible with itself is duplicated in tandem, and the connected segments diverge to accommodate new functions within the tolerance of structural compatibility [2]. Splicing and duplication of genes govern the highly efficient and economic ways — by which higher order structures are built from recycling and repeating basic modules — to engineer structurally coupled functions. As a consequence, the repeated segments that make up a repeat protein are often not identical. However, the modularity inspires engineering efforts. In this review, we discuss the methods developed for designing repeat proteins, their achievements, and the new directions ahead.

## Repeat protein architecture

The repeated modularity implies that there is a continuum of structural features from section to section so that each unit can attach to a preceding one while maintaining the features for a subsequent unit. Most of these units fall in the range between 5 and 50 residues and fold into a domain only in the presence of the neighboring units [3]. These structures can be very similar to other globular proteins if only a small number of repeating units are the constituents, but can also be uniquely non-globular as elongated rods, arches, horseshoes, as increasingly large number of repeating units are involved. The secondary structure elements of a repeat protein form a base coil with either right- or left-handed spiral handedness, whereas the overall structure of the protein twists further into a super-helix [4] (Fig.1a and 1b). The parameters that describe a superhelix (radius, twist, and rise) can be

extracted by fitting equations over the geometric centers of the modular units. These three parameters can describe a wide range of structures, and for native repeat families, their fitted parameters distribute over a narrow, characteristic range [5]. In all cases each repeat interacts primarily only with the adjacent units. The conserved features for a repeat protein family are usually the hydrophobic residues through which the modular units interact and the few key topologically relevant residues responsible for defining each unique shape [FIG 1c and 1d]. For open ended non-globular structures, however, the terminal capping repeats show remarkable sequence difference. Capping repeats are responsible for shielding the extended hydrophobic core shared among repeats from solvent. These caps are important for solubility and are nucleation sites required for some repeat proteins to fold [6]. In the past 15 years, extensive engineering efforts have been undertaken to design and study proteins with modular features, particularly the types that can fold cooperatively into single domains [7].

Designing repeat proteins

Repeat proteins display unique characteristics that make them attractive to design. The similarity between the repeating units within a protein on both structure and sequence levels suggests the possibility of designing an idealized unit with characteristics of the family. Repeat protein's limited interactions to only neighboring repeats also presents significantly less complexity compared to designing globular proteins where long-range contacts are abundant and often irregular. Repeat units can be inserted, removed or replaced without significantly affecting the overall structure as long as the compatible interfaces between units are present [8]. Indeed, the intended goal of designing repeat proteins is to create modular systems from which custom scaffolds can be created for various applications [7]. Most engineering efforts on repeat proteins to date are on the design of idealized versions of naturally occurring repeat units that maintain stable, homogeneous and modular structures (Fig.2a). Modularity can be explored on two different levels: sequence analysis of native protein motifs and structure based design. The goal is to create a variety of modules to generate custom shapes. (Fig 2b)

Design from sequence

Since repeat protein families are defined by their sequence patterns, the information about residue conservation can be extracted from a multiple sequence alignment (MSA). The abundance of evolutionary related sequences, including the repeating units within the same protein, allows for an effective and widespread use of MSA information. The sequence with the most frequent amino acid at each position is the "consensus". The design process involves collecting sequences of repeat units, building an MSA and deriving the consensus as the new idealized repeat sequence (Fig.2c). Multiple copies of the unit are then linked to create a homogeneous repeat protein sequence; capping repeats are derived from naturally existing proteins. Consensus design has been the first and most widely used approach across multiple repeat protein families: ankyrin (ANK) [9–13], tetratricopeptide repeat (TPR) [14], armadillo (ARM) [15–17], leucine rich repeat (LRR) [18], pentatricopeptide repeat (PPR) [19,20], pentapeptide repeat [21], 42 residue TPR variant (42PR) [22] and HEAT-EZ [23], all have designed structures confirmed by crystallography (Table 1). Effective consensus design relies on the assumption that evolutionary conservation translates into self-compatibility of repeats and foldability of the structures. This assumption holds true for

highly conserved families like ANK and TPR, but becomes problematic for other families with greater repeat variation, like the ARM family [24]. Low sequence similarity, short sequence length and problematic detection of tandem repeats with high confidence limit automated generation of accurate MSAs [3]. Further restricting the MSA to narrow and homogeneous groups within the families produced successful designs in some cases [18,23]. Using structure information and molecular dynamic simulations produced stable ARM designs; ARM designs from consensus were molten globule-like [15,16]. Instead of extracting a consensus sequence, Lee and Blaber [25,26] progressively introduced mutations experimentally to a beta trefoil background to arrive at sequences that contain three identical tandem repeats. Recently Smock *et al.* [27] analyzed the formation of a tachylectin precursor through design and selection of libraries of potential ancestors with identical repeats, inferring phylogenetic distance and progenitor sequences from MSA. However, when there is no clear detectable signal or motif in the sequence, sequence based methods are not likely to succeed without guidance from structural information.

## Structure based design

### Designs using native structural templates

Consensus sequences from MSAs are often threaded onto known structures for validation. A more extensive use of structural information involves using known structures as templates for deriving sequences. Structural templates constraint the types of residues compatible with the protein architecture, which should directly reflect the conserved amino acid features on structural basis. Templates used can range from a single repeat unit, to multiple units, to a full repeat protein domain. When a native template with highly symmetric backbone is available, designs can be carried out directly on the full structure, as shown by Broom *et al.* [28], who generated a symmetric beta trefoil by applying a consensus sequence in the most conserved positions of a starting template and designing the rest of the sequence with the Rosetta molecular modeling suite [29].

Ideally for structure based design, a single repeat template unit should allow arbitrary propagation to a full length structure. The challenge, however, is with modeling properly the junctions connecting the units. The concept of propagating a single unit is shown possible in three cases using Rosetta to resolve issues in the connecting junctions and design the sequence. Zhu *et al.* [30] investigated the evolution of tandem repeat proteins by designing a novel TPR from a ribosomal protein with high structural similarity to a single repeat. A designed TPR [14] was used as superposition template to determine the relative orientation of the repeats and as source for the connecting loop. Voet *et al.* [31] generated a series of beta propellers from the NHL family, deriving from a single repeat backbone that was symmetrically docked. The repeat backbones were connected by a glycine residue and sequences were designed through ancestral reconstruction. Rämisch *et al.* [32] employed also a computational docking strategy to select LRRs suitable to generate a novel toroid structure. Here they rebuilt the connecting loops by modeling a threonine residue and designed the repetitive sequence [33] (Fig.2d). Template based design has been used on a smaller number of families than consensus design because only a small portion of templates have sufficient structural similarity between repeats.

Template-free (*de novo*) design

Computational protein design methods [34] allow generation of repeat protein backbones with desired features and perfect internal symmetry through controlling the torsional angles of the peptide chain. The ability to model structures at this level implies that repeat proteins of any length can be built with their sequences designed simultaneously across repeats. Backbone design proceeds through a definition of the desired secondary structure, usually that of a single repeating unit. The internal repeat symmetry then mirrors the behavior of the base unit to all other synchronized copies (Fig.2e). Sequence design is then performed similarly on all of the copies simultaneously. As *de novo* design allows complete control of the structure, idealized units of native repeat families can also be directly created by incorporating native features as sequence and structure constraints. This approach readily created idealized ANK, TPR, LRR, WD40, HEAT and ARM repeat proteins [35,36]. With the exception of LRRs, capping repeats could be directly designed from the structures.

Evolutionary information is not always necessary for computational design. A *de novo* four-fold symmetric triosephosphate isomerase (TIM) barrel based solely on barrel geometry considerations [37] presented a simplified structural view of the vastly diverse protein family, which has eluded design efforts for decades. The highly cooperative and symmetrical structure allowed annotation of fold determinant features independent of evolutionary influence. While the *de novo* TIM barrel is a novel protein with no close native homologs, the $(\beta/\alpha)_8$ barrel is still a native fold. Going further, the constraints associated with repeat proteins, namely the internal symmetry and the very simple secondary structure organization within a unit, make repeat proteins a platform for addressing the fundamental question of whether nature has covered most of the fold space. Although there is a predominance of right handed (on the base coil level) α helical repeats in nature, Doyle et *al.* [38] designed left-handed α toroids with novel topologies (designed helical toroids, DHT). Brunette *et al.* [5] automated a method for sampling α solenoids, and the space it sampled not only covered common native geometries such as TPRs, but also a large number of equally designable right- and left-handed repeats of very distinct shapes apart from known repeat structures (designed helical repeats, DHR). All these results support the idea that the protein fold space may remain largely unexplored by evolution [34].

Outlook on repeat protein design

The body of work in the past 15 years has consolidated design methods for repeat proteins, but recently new directions have emerged:

**Parametric design.** Using parametric equations to systematically guide the exploration of backbone conformations has achieved unprecedented stabilities and oligomeric specificities in designed coiled coils [39–41]. It is likely that parametric equations can be derived for repeat proteins — instead of describing the general periodicity in α helices, the equations can represent transformations of repeat units in the same fashion (Fig 1b).

**Design of intrinsically disordered repeat proteins (IDRPs).** A large number of naturally occurring IDRPs exists [42] and have recently spurred much research interest. Quiroz and Chilkoti [43] designed IDRPs to systematically study the sequence encoded phase behavior. IDRP characterization has relied on distinctive characteristics other than conventional biophysics methods for folded proteins; this area is still largely unexplored by protein design.

**Modular control.** Creating customized structures by combining modular units remains a challenge, but we started to see the potential for designing this kind of repeat proteins. Using contiguous and discontinuous motifs, Jacobs *et al.* [44] developed a superposition based design method that could potentially be used to build new repeat proteins. Park *et al.* [45] have shown for the first time that it is possible to combine different repeats (all LRRs in this case) in a single domain. Because the different types of LRR repeats natively encode different overall curvature, combining varying numbers of modules in different order creates controllable shapes (Fig 2b). Furthermore, repeat proteins have been designed into oligomers [46,47] and split repeat protein systems [48] (Fig 3), and it is likely that all these modules can be further combined in the future to generate novel proteins using different types of repeats and quaternary structure.

The past years have seen an accelerated development of new repeat proteins and tools for their assembly. With growing collection of modular parts amassed and new technology developed, turning repeat proteins into new materials in biotechnology becomes a real possibility. The new frontier ahead promises development of functionalized platforms for very diverse applications.

FIGURE LEGENDS

Figure 1. General features of repeat proteins. (A) Example of a HEAT repeat family member (PDB: 3dw8). (B) The parametric description of the structure in (A). Repeat units stack against one another into a base coil, which generally has a uniform curvature that can be described by a super helix. (C) Structural features of an ARM repeat family member (PDB: 3nmw). An ARM repeat unit has 3 α helices; a very distinct pattern of glycines (yellow residues shown in spheres) is the basis for transitioning between two of the helices. The pink stripe in the central repeat shows positions that are 3 residues apart along the sequence. The residues should mark the same side of a helix if it were straight, but the glycine feature induces bending and rotation of the helices, resulting in the residues on the pink stripe located on opposite sides of the helices. A dash line is shown connecting the equivalent positions on different units. This represents the twist along the base coil, and it can be different for different repeat families. (D) The structural features of an ANK repeat family (PDB: 4hi8). A highly conserved “TPLH” motif along with other polar residues in the region (yellow stick residues) are responsible for stabilizing the characteristic hairpins associated with ANK repeat units. ANK has a very different base coil level twist from ARM repeats.

Figure 2. Designing repeat proteins. (A) A general goal of designing repeat proteins is to create modular units that can be expanded into functionalizing scaffolds. (B) As long as the modules are compatible with one another, combining different types of units allows control in overall structure. (C) Sequence based design. From the alignment of evolutionarily related repeat units (shown as a logo plot, weblogo.berkeley.edu), a consensus (bottom row sequence) can be derived from the most common amino acids at each position. (D) Template-based design. The general method developed by Rämisch *et al.* [32] identifies a

unit within a repeat protein (in this case, a leucine rich repeat of porcine ribonuclease inhibitor; PDB: 2bnh) and uses symmetric docking, backbone rebuilding and sidechain optimization methods to achieve the design. They modeled the horseshoe-shaped native LRR into a toroid with C10 symmetry and showed experimentally that half circle constructs can self assemble into the full ring. (E) *De novo* design. Backbones are built directly from secondary structure elements and are not borrowed from known native structures. The units (colored by different colors: red, yellow, blue green) are identical and are simulated in a synchronized fashion to maintain an internal repeat symmetry on both backbone and sidechain levels.

Figure 3. Emergent self-assembling property of repeat proteins. Idealized repeat proteins that can fold without capping repeats can often self-assemble into oligomers. An idealized ARM repeat was shown to form dimers in solution. Most of the other reported cases are from toroid designs. The designs were made as toroids, but when a fraction of the repeat protein was expressed, oligomers were observed in solution. Half length ribonuclease inhibitor-type LRR and half length TIM barrel form dimers, and a DHT assemble into a tetramer. Monomer units are shown in color; the partners are in grey. For the DHT tetramer, the three bind partners are in different shades of grey.

References

• of special interest
•• of outstanding interest


1. Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D: **A census of protein repeats1**. *J. Mol. Biol.* 1999, **293**:151–160.
2. Andrade MA, Perez-Iratxeta C, Ponting CP: **Protein Repeats: Structures, Functions, and Evolution**. *J. Struct. Biol.* 2001, **134**:117–131.
3. Kajava AV: **Tandem repeats in proteins: From sequence to structure**. *J. Struct. Biol.* 2012, **179**:279–288.
4. Kobe B, Kajava AV: **When protein folding is simplified to protein coiling: the continuum of solenoid protein structures**. *Trends Biochem. Sci.* 2000, **25**:509–515.
••5. Brunette TJ, Parmeggiani F, Huang P-S, Bhabha G, Ekiert DC, Tsutakawa SE, Hura GL, Tainer JA, Baker D: **Exploring the repeat protein universe through computational protein design**. *Nature* 2015, **528**:580–584.
6. Courtemanche N, Barrick D: **The Leucine-Rich Repeat Domain of Internalin B Folds along a Polarized N-Terminal Pathway**. *Structure* 2008, **16**:705–714.
7. Javadi Y, Itzhaki LS: **Tandem-repeat proteins: regularity plus modularity equals design-ability**. *Curr. Opin. Struct. Biol.* 2013, **23**:622–631.
8. Tripp KW, Barrick D: **The Tolerance of a Modular Protein to Duplication and Deletion of Internal Repeats**. *J. Mol. Biol.* 2004, **344**:169–178.
9. Mosavi LK, Minor DL, Peng Z: **Consensus-derived structural determinants of the ankyrin repeat motif**. *Proc. Natl. Acad. Sci.* 2002, **99**:16029–16034.

10. Binz HK, Stumpp MT, Forrer P, Amstutz P, Plückthun A: **Designing Repeat Proteins: Well-expressed, Soluble and Stable Proteins from Combinatorial Libraries of Consensus Ankyrin Repeat Proteins**. *J. Mol. Biol.* 2003, **332**:489–503.
11. Merz T, Wetzel SK, Firbank S, Plückthun A, Grütter MG, Mittl PRE: **Stabilizing Ionic Interactions in a Full-consensus Ankyrin Repeat Protein**. *J. Mol. Biol.* 2008, **376**:232–240.
12. Tripp KW, Barrick D: **Enhancing the Stability and Folding Rate of a Repeat Protein through the Addition of Consensus Repeats**. *J. Mol. Biol.* 2007, **365**:1187–1200.
13. Aksel T, Majumdar A, Barrick D: **The Contribution of Entropy, Enthalpy, and Hydrophobic Desolvation to Cooperativity in Repeat-Protein Folding**. *Structure* 2011, **19**:349–360.
14. Main ERG, Xiong Y, Cocco MJ, D'Andrea L, Regan L: **Design of Stable α-Helical Arrays from an Idealized TPR Motif**. *Structure* 2003, **11**:497–508.
15. Parmeggiani F, Pellarin R, Larsen AP, Varadamsetty G, Stumpp MT, Zerbe O, Caflisch A, Plückthun A: **Designed Armadillo Repeat Proteins as General Peptide-Binding Scaffolds: Consensus Design and Computational Optimization of the Hydrophobic Core**. *J. Mol. Biol.* 2008, **376**:1282–1304.
16. Alfarano P, Varadamsetty G, Ewald C, Parmeggiani F, Pellarin R, Zerbe O, Plückthun A, Caflisch A: **Optimization of designed armadillo repeat proteins by molecular dynamics simulations and NMR spectroscopy**. *Protein Sci.* 2012, **21**:1298–1314.
17. Reichen C, Madhurantakam C, Plückthun A, Mittl PRE: **Crystal structures of designed armadillo repeat proteins: Implications of construct design and crystallization conditions on overall structure**. *Protein Sci.* 2014, **23**:1572–1583.
18. Lee S-C, Park K, Han J, Lee J, Kim HJ, Hong S, Heu W, Kim YJ, Ha J-S, Lee S-G, et al.: **Design of a binding scaffold based on variable lymphocyte receptors of jawless vertebrates by module engineering**. *Proc. Natl. Acad. Sci.* 2012, **109**:3299–3304.
19. Coquille S, Filipovska A, Chia T, Rajappa L, Lingford JP, Razif MFM, Thore S, Rackham O: **An artificial PPR scaffold for programmable RNA recognition**. *Nat. Commun.* 2014, **5**:5729.
20. Gully BS, Shah KR, Lee M, Shearston K, Smith NM, Sadowska A, Blythe AJ, Bernath-Levin K, Stanley WA, Small ID, et al.: **The design and structural characterization of a synthetic pentatricopeptide repeat protein**. *Acta Crystallogr. D Biol. Crystallogr.* 2015, **71**:196–208.
21. MacDonald JT, Kabasakal BV, Godding D, Kraatz S, Henderson L, Barber J, Freemont PS, Murray JW: **Synthetic beta-solenoid proteins with the fragment-free computational design of a beta-hairpin extension**. *Proc. Natl. Acad. Sci.* 2016, **113**:10346–10351.
22. Marold JD, Kavran JM, Bowman GD, Barrick D: **A Naturally Occurring Repeat Protein with High Internal Sequence Identity Defines a New Class of TPR-like Proteins**. *Structure* 2015, **23**:2055–2065.
23. Urvoas A, Guellouz A, Valerio-Lepiniec M, Graille M, Durand D, Desravines DC, van Tilbeurgh H, Desmadril M, Minard P: **Design, Production and Molecular Structure of a**

**New Family of Artificial Alpha-helicoidal Repeat Proteins (αRep) Based on Thermostable HEAT-like Repeats**. *J. Mol. Biol.* 2010, **404**:307–327.
24. Sawyer N, Chen J, Regan L: **All Repeats Are Not Equal: A Module-Based Approach to Guide Repeat Protein Design**. *J. Mol. Biol.* 2013, **425**:1826–1838.
25. Lee J, Blaber SI, Dubey VK, Blaber M: **A Polypeptide "Building Block" for the β-Trefoil Fold Identified by "Top-Down Symmetric Deconstruction."** *J. Mol. Biol.* 2011, **407**:744–763.
26. Lee J, Blaber M: **Experimental support for the evolution of symmetric protein architecture from a simple peptide motif**. *Proc. Natl. Acad. Sci.* 2011, **108**:126–130.
27. Smock RG, Yadid I, Dym O, Clarke J, Tawfik DS: **De Novo Evolutionary Emergence of a Symmetrical Protein Is Shaped by Folding Constraints**. *Cell* 2016, **164**:476–486.
28. Broom A, Doxey AC, Lobsanov YD, Berthin LG, Rose DR, Howell PL, McConkey BJ, Meiering EM: **Modular Evolution and the Origins of Symmetry: Reconstruction of a Three-Fold Symmetric Globular Protein**. *Structure* 2012, **20**:161–171.
29. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman K, Renfrew PD, Smith CA, Sheffler W, et al.: **ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules**. *Methods Enzymol.* 2011, **487**:545–574.
30. Zhu H, Sepulveda E, Hartmann MD, Kogenaru M, Ursinus A, Sulz E, Albrecht R, Coles M, Martin J, Lupas AN: **Origin of a folded repeat protein from an intrinsically disordered ancestor**. *eLife* 2016, **5**:e16761.
•31. Voet ARD, Noguchi H, Addy C, Simoncini D, Terada D, Unzai S, Park S-Y, Zhang KYJ, Tame JRH: **Computational design of a self-assembling symmetrical β-propeller protein**. *Proc. Natl. Acad. Sci.* 2014, **111**:15102–15107.
•32. Rämisch S, Weininger U, Martinsson J, Akke M, André I: **Computational design of a leucine-rich repeat protein with a predefined geometry**. *Proc. Natl. Acad. Sci.* 2014, **111**:17875–17880.
33. DiMaio F, Leaver-Fay A, Bradley P, Baker D, André I: **Modeling Symmetric Macromolecular Structures in Rosetta3**. *PLoS ONE* 2011, **6**:e20450.
•34. Huang P-S, Boyken SE, Baker D: **The coming of age of de novo protein design**. *Nature* 2016, **537**:320–327.
•35. Parmeggiani F, Huang P-S, Vorobiev S, Xiao R, Park K, Caprari S, Su M, Seetharaman J, Mao L, Janjua H, et al.: **A General Computational Approach for Repeat Protein Design**. *J. Mol. Biol.* 2015, **427**:563–575.
36. Reichen C, Hansen S, Forzani C, Honegger A, Fleishman SJ, Zhou T, Parmeggiani F, Ernst P, Madhurantakam C, Ewald C, et al.: **Computationally Designed Armadillo Repeat Proteins for Modular Peptide Recognition**. *J. Mol. Biol.* 2016, **428**:4467–4489.
••37. Huang P-S, Feldmeier K, Parmeggiani F, Fernandez Velasco DA, Höcker B, Baker D: **De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy**. *Nat. Chem. Biol.* 2016, **12**:29–34.
••38. Doyle L, Hallinan J, Bolduc J, Parmeggiani F, Baker D, Stoddard BL, Bradley P: **Rational design of α-helical tandem repeat proteins with closed architectures**. *Nature* 2015, **528**:585–588.

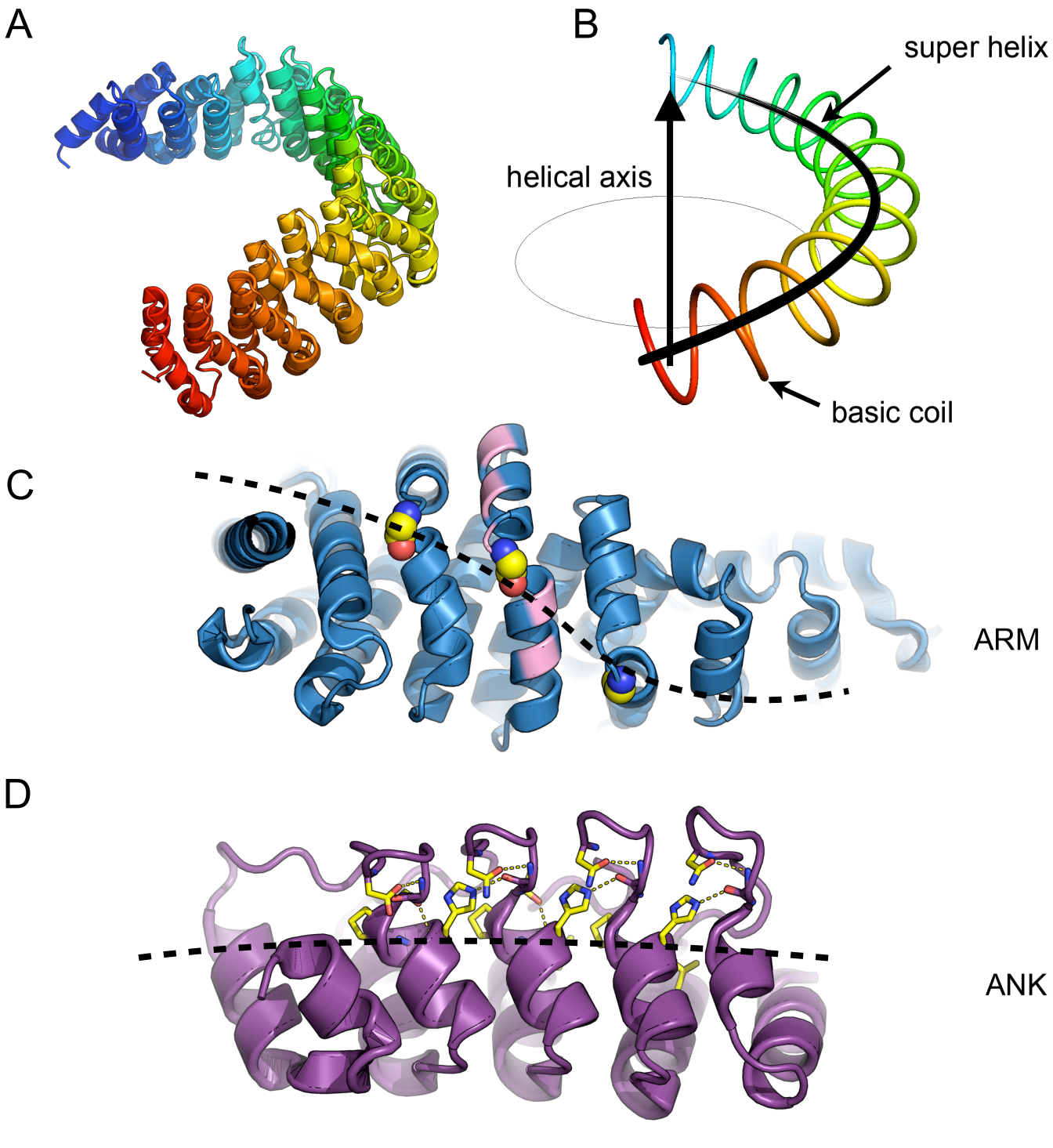
39. Huang P-S, Oberdorfer G, Xu C, Pei XY, Nannenga BL, Rogers JM, DiMaio F, Gonen T, Luisi B, Baker D: **High thermodynamic stability of parametrically designed helical bundles**. *Science* 2014, **346**:481–485.
40. Thomson AR, Wood CW, Burton AJ, Bartlett GJ, Sessions RB, Brady RL, Woolfson DN: **Computational design of water-soluble α-helical barrels**. *Science* 2014, **346**:485–488.
41. Boyken SE, Chen Z, Groves B, Langan RA, Oberdorfer G, Ford A, Gilmore JM, Xu C, DiMaio F, Pereira JH, et al.: **De novo design of protein homo-oligomers with modular hydrogen-bond network–mediated specificity**. *Science* 2016, **352**:680–687.
42. Jorda J, Xue B, Uversky VN, Kajava AV: **Protein tandem repeats: the more perfect the less structured**. *Febs J.* 2010, **277**:2673–2682.
•43. Quiroz FG, Chilkoti A: **Sequence heuristics to encode phase behaviour in intrinsically disordered protein polymers [Internet]**. *Nat. Mater.* 2015, **advance online publication**.
•44. Jacobs TM, Williams B, Williams T, Xu X, Eletsky A, Federizon JF, Szyperski T, Kuhlman B: **Design of structurally distinct proteins using strategies inspired by evolution**. *Science* 2016, **352**:687–690.
••45. Park K, Shen BW, Parmeggiani F, Huang P-S, Stoddard BL, Baker D: **Control of repeat-protein curvature by computational protein design**. *Nat. Struct. Mol. Biol.* 2015, **22**:167–174.
•46. Fallas JA, Ueda G, Sheffler W, Nguyen V, McNamara DE, Sankaran B, Pereira JH, Parmeggiani F, Brunette TJ, Cascio D, et al.: **Computational design of self-assembling cyclic protein homo-oligomers [Internet]**. *Nat. Chem.* 2016, doi:10.1038/nchem.2673.
47. Mills JH, Sheffler W, Ener ME, Almhjell PJ, Oberdorfer G, Pereira JH, Parmeggiani F, Sankaran B, Zwart PH, Baker D: **Computational design of a homotrimeric metalloprotein with a trisbipyridyl core**. *Proc. Natl. Acad. Sci.* 2016, **113**:15012–15017.
48. Watson RP, Christen MT, Ewald C, Bumbak F, Reichen C, Mihajlovic M, Schmidt E, Güntert P, Caflisch A, Plückthun A, et al.: **Spontaneous Self-Assembly of Engineered Armadillo Repeat Protein Fragments into a Folded Structure**. *Structure* 2014, **22**:985–995.

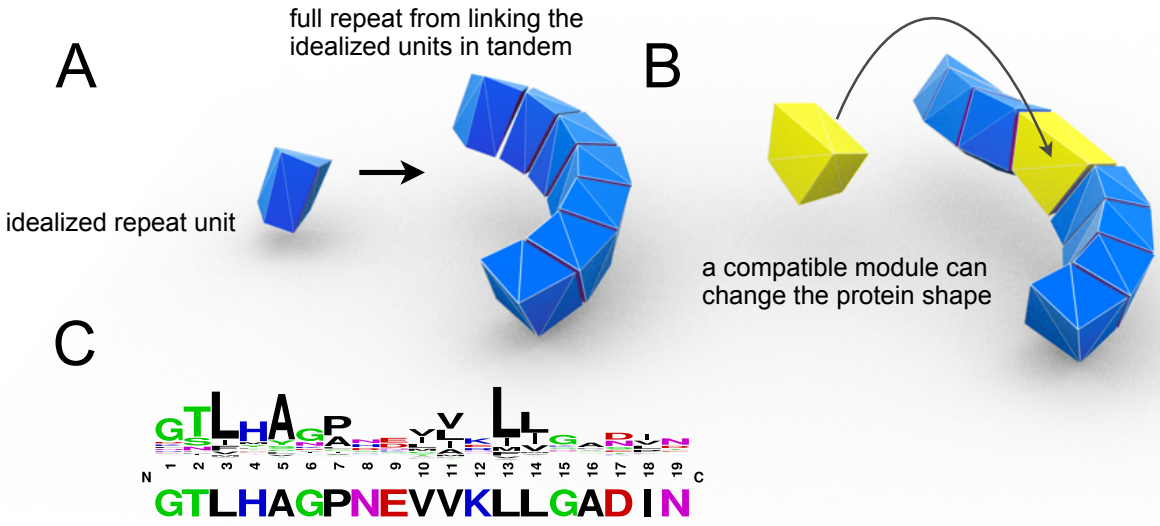
Table 1. Designed repeat proteins with representative structures

| Repeat protein | Type of fold | Repeat length | Method | Structures and references |
|---|---|---|---|---|
| 42PR | α solenoid | 42 | Consensus design | 4Y6C, 4Y6W [22] |
| ankyrin | α solenoid | 33 | Consensus design<br>Consensus design<br>Consensus design<br>*De novo* backbone design | 1N0Q, 1N0R [9]<br>2QYJ [10,11]<br>2L6B [12,13]<br>4HQD, 4HB5, 4GPM, 4GMR [35] |
| armadillo | α solenoid | 42 | Consensus design<br>*De novo* backbone design<br>*De novo* backbone design | 4PLS, 4PLQ, 4PLR [15–17]<br>4HXT, 4RV1, 4RZP [35]<br>4D49, 4D4E [36] |
| β trefoil | β toroid | 52<br>47 | Rational design, mutagenesis<br>Template structure, ancestral reconstruction | 3O49, 3O4A, 3O4B, 3O4C, 3O4D [25,26]<br>3PG0 [28] |
| DHR | α solenoid | 37-61 | *De novo* backbone design | 5CWB, 5CWC, 5CWD, 5CWF, 5CWG, 5CWH, 5CWI, 5CWJ, 5CWK, 5CWL, 5CWM, 5CWN, 5CWO, 5CWP, 5CWQ [5] |
| DHT | α toroid | 31-35 | *De novo* backbone design | 4YXX, 4YXY, 4YXZ, 4YY2, 4YY5, 5BYO [38] |
| HEAT-EZ | α solenoid | 31 | Consensus design | 3LTJ, 3LTM [23] |
| LRR | α/β solenoid | 24<br>24 | Consensus design | 3RFJ, 3RFS [18]<br>4PSJ, 4PQ8 [35] |

| | | | | |
|---|---|---|---|---|
| | | 22,24,28+29 | *De novo* backbone design<br>*De novo* backbone design | 4R58, 4R5C, 4R5D, 4R6F, 4R6G, 4R6J [45] |
| NHL | β propeller | 43 | Template structure, ancestral reconstruction | 3WW7, 3WW8, 3WW9, 3WWa, 3WWB, 3WWF [31] |
| Pentapeptide | β helix | 5 | Consensus design | 4YFO, 4YC5 [21] |
| PPR | α solenoid | 35 | Consensus design<br>Consensus design | 4PJQ [19]<br>4WN4, 4WSL, 4OZS [20] |
| Tachylectin | β propeller | 47 | Ancestral reconstruction, selection | 5C2N* [27] |
| TIM barrel | α/β barrel | 46 | *De novo* backbone design | 5BVL [37] |
| TPR | α solenoid | 34 | Consensus design<br>Template structure, ancestral reconstruction | 1NA0, 1NA3 [14]<br>5FZQ, 5FZR, 5FZS [30] |

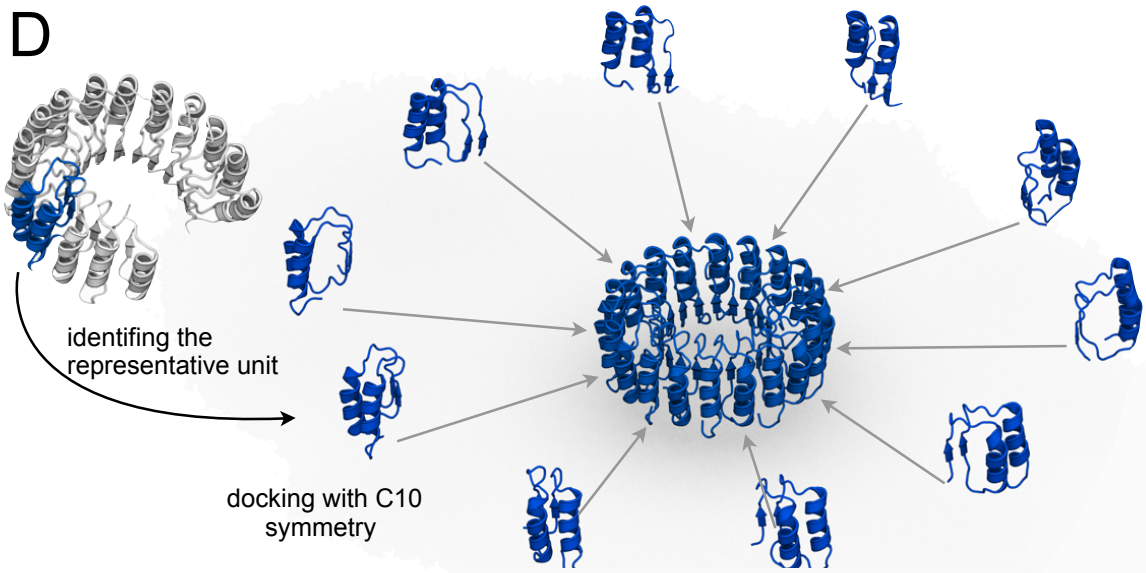
* crystal structure of the oligomeric precursor of the repeat protein

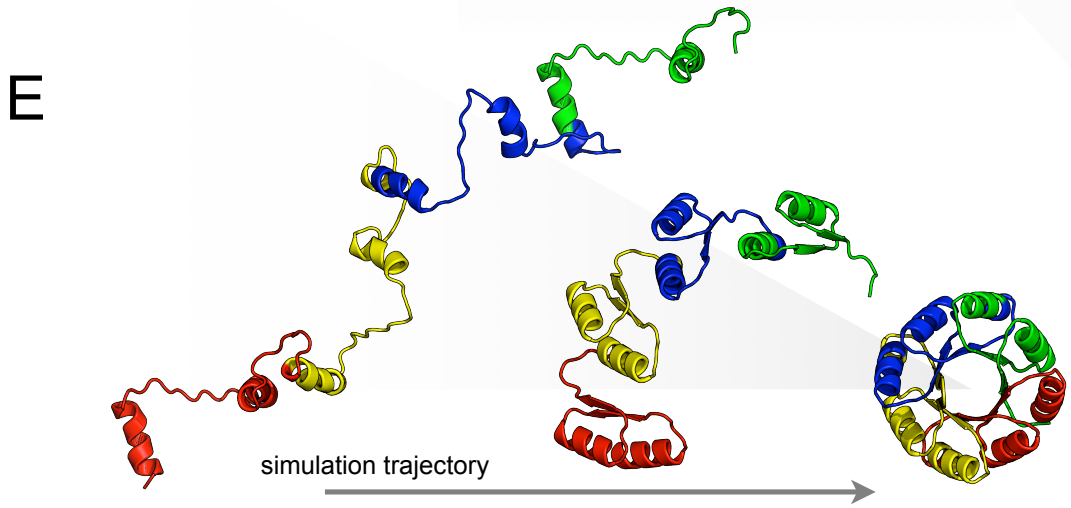

A
B
super helix
helical axis
basic coil
C
ARM
D
ANK

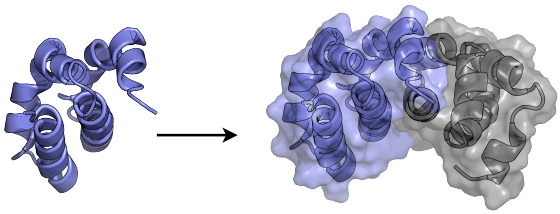

A
full repeat from linking the
idealized units in tandem
idealized repeat unit
B
a compatible module can
change the protein shape

C

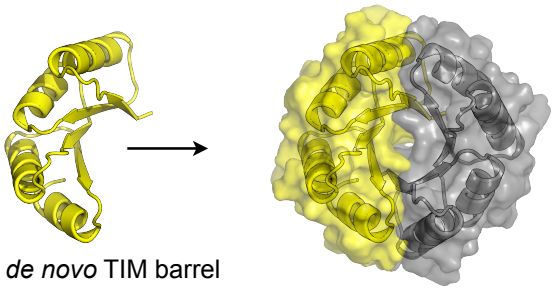

N
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
C
GTLHAGPNEVVKLLGADIN


D
identifing the
representative unit
docking with C10
symmetry

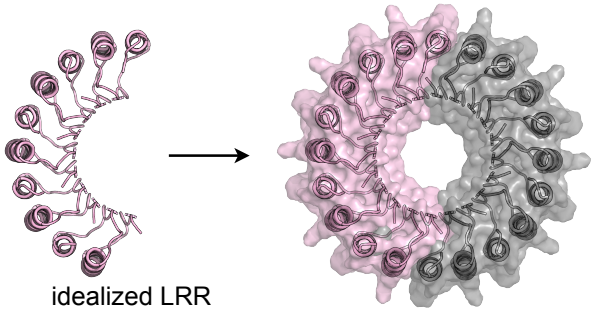

E
simulation trajectory



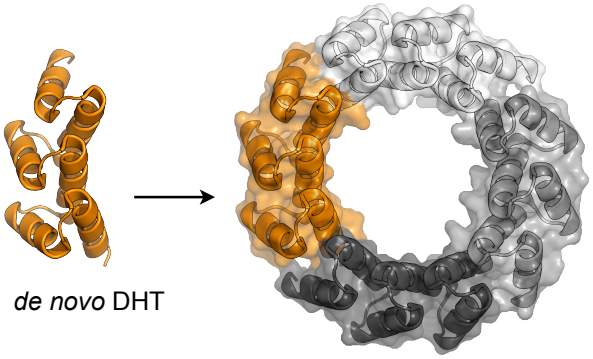
idealized ARM



*de novo* TIM barrel



idealized LRR



*de novo* DHT